Class 2: Kalman Filter

Macroeconometrics - Spring 2015

Jacek Suda

March 23, 2015

Outline

Outline:

1. Estimation
   - OLS
   - MLE
   - AR(1)
   - MA(1)
2. Forecasting
3. Prediction Error Decomposition
4. State-Space Form
5. Kalman Filter

ARMA

- Wold representation:

$$Y_t = \kappa_t + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \qquad \varepsilon_t \sim WN(0, \sigma^2), \qquad \sum_{j=0}^{\infty} \psi_j^2 < \infty$$

- AR(p):

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t, \qquad \varepsilon_t \sim WN$$

- MA(q):

$$Y_t - \mu = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \qquad \varepsilon_t \sim WN$$

- ARMA(p,q):

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

ESTIMATION

Estimation

- OLS
- MLE
- AR(1)
- MA(1)

OLS

For AR(p), OLS is equivalent to Conditional MLE

- Model:

$$\begin{aligned} y_t &= c + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2). \\ y_t &= x_t'\beta + \varepsilon_t, \\ & \beta = (c, \phi_1, \phi_2, \ldots, \phi_p), \qquad x_t = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \end{aligned}$$

- OLS:

$$\begin{aligned} \hat{\beta} &= \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t y_t, \\ \hat{\sigma}^2 &= \frac{1}{T - (p+1)} \sum_{t=1}^{T} (y_t - x_t'\hat{\beta})^2. \end{aligned}$$

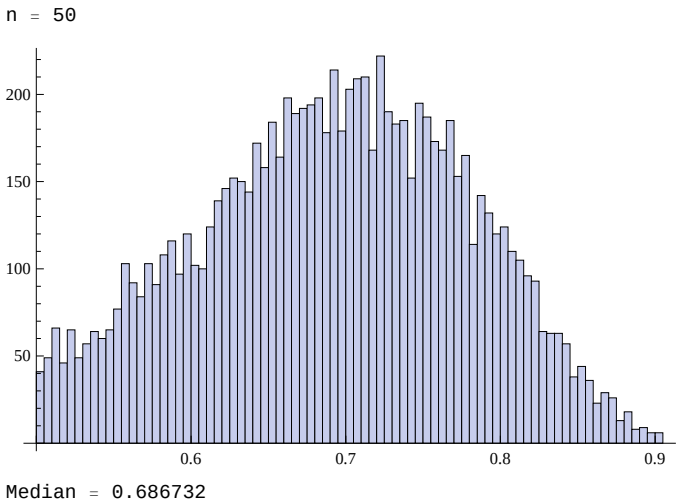
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



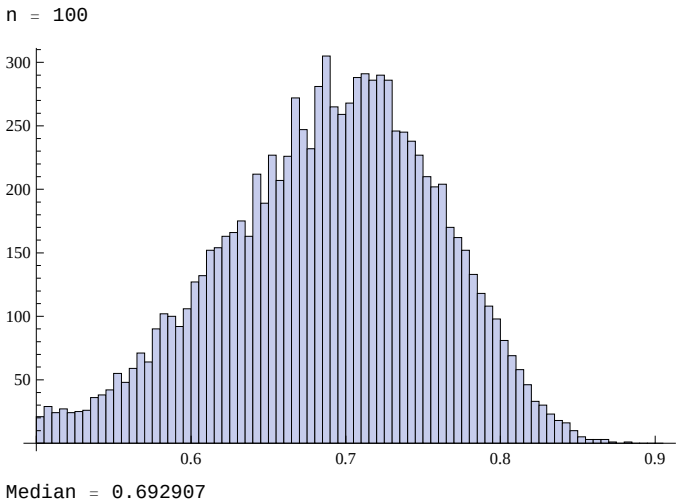
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



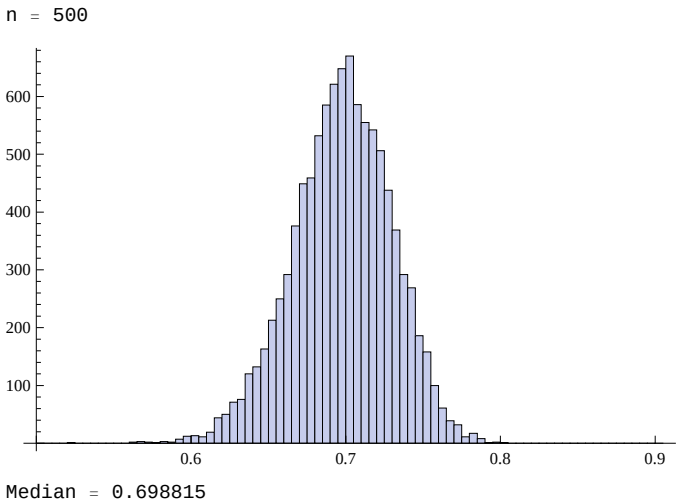
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



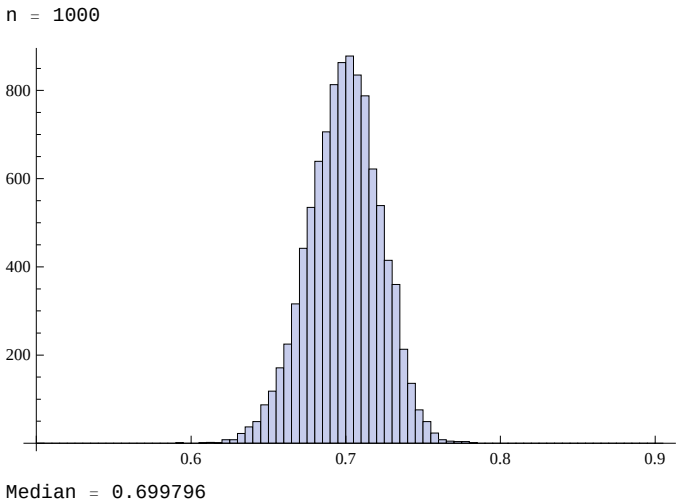
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



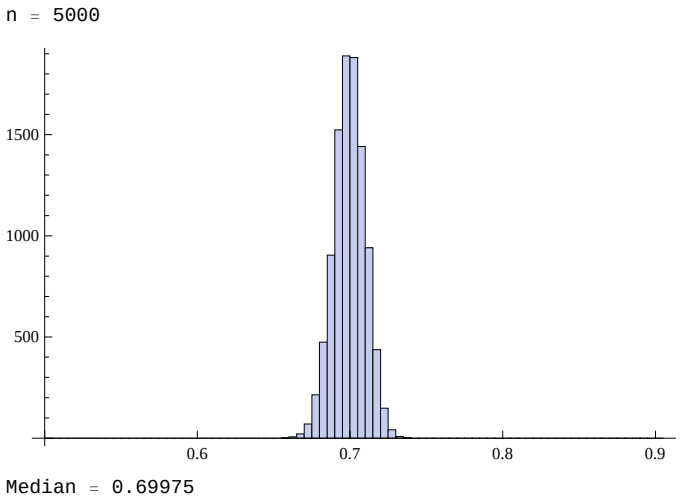
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



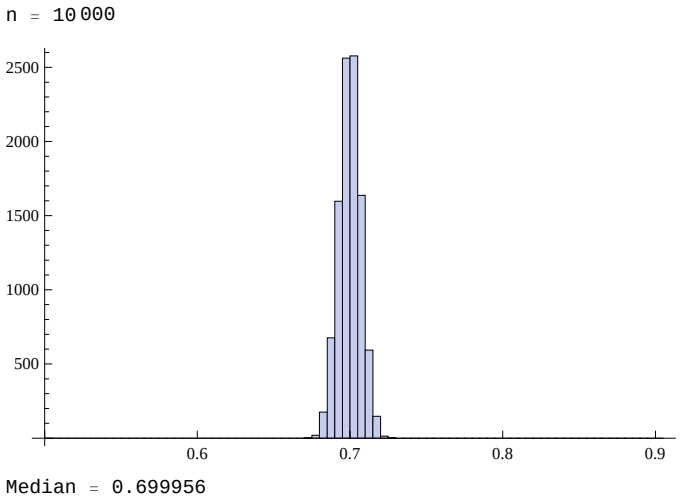
Properties

Note:

- $E[\hat{\beta}] \neq \beta$ because x_t is random and $E(\varepsilon|Y) \neq 0$. But, if $|z| > 1$ for $\phi(z) = 0$ (or $|\lambda| < 1$ for F) then,

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

- *Estimator might be biased but consistent, it converges in probability.*
- Illustration:
  $\phi = 0.7$



Hypothesis testing

Hypothesis testing:

$$\sqrt{T}(\hat{\beta}-\beta) \xrightarrow{d} N(0,\sigma^2 V^{-1}), \qquad V = plim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} x_t x_t'.$$

- *If we have enough data* $(T \to \infty)$ *then t-student will converge to Normal distribution.*

$$t^{\beta=\beta_0} = \frac{\hat{\beta}-\beta_0}{\hat{SE}(\hat{\beta})} \sim t-student(T-1) \underset{T\to\infty}{\longrightarrow} N(0,1).$$

- *See Hamilton for a downward bias in a finite sample, i.e.* $E[\hat{\beta}] < \beta$.

MLE

If $Y_t \sim iid$, joint likelihood is the product of marginal likelihoods (pdf)

$$\begin{aligned} L(\tilde{\theta}|y_1,\ldots,y_T) &= \prod_{t=1}^{T} L(\tilde{\theta}|y_t), \\ &= \prod_{t=1}^{T} f(y_t|\tilde{\theta}), \end{aligned}$$

Different parameters have different probability of generating $\{y_1,\ldots,y_T\}$. But if Y_t is not independent, factorization is invalid. Instead, factor into conditional distributions.

$$\begin{aligned} f(Y_1,Y_2|\tilde{\theta}) &= f(Y_2|Y_1,\tilde{\theta})f(Y_1|\tilde{\theta}), \\ f(Y_1,Y_2,Y_3|\tilde{\theta}) &= f(Y_3|Y_2,Y_1,\tilde{\theta})f(Y_2,Y_1|\tilde{\theta}), \\ L(\tilde{\theta}|y_1,\ldots,y_T) = f(Y_1,\ldots,Y_T|\tilde{\theta}) &= \prod_{t=2}^{T} f(Y_t|Y_{t-1},\ldots,Y_2,Y_1,\tilde{\theta})f(Y_1|\tilde{\theta}) \end{aligned}$$

MLE

- Conditional MLE assumes Y_1 fixed (not random). *It's just a simplifying assumption.*

$$L^c(\tilde{\theta}|y_1,\ldots,y_T) = \prod_{t=2}^{T} f(Y_t|Y_{t-1},\ldots,Y_1,\tilde{\theta}).$$

- Given normality, first order conditions of maximization L^c are linear in $\tilde{\theta}$.
- For AR models $\hat{\phi}^{CMLE} \Leftrightarrow OLS$.
- Conditional MLE is consistent. (*It's not so efficient as we ignore randomness of* Y_1.)
- Exact MLE requires non-linear optimization.
- *Often we don't know distribution of the the data. One thing that is assumed in OLS is the* $\varepsilon \sim WN$.
- *Assuming normality in CMLE is not a bad assumption as we still get consistent estimator: quasi MLE. (see Davidson and MacKinnon.)*
- *We don't get unbiasness in any of MLE. What can be done is to compute what the bias is and correct for it.*

Estimation AR(1)

Recall: AR(1)

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \qquad \varepsilon \sim iidN(0, \sigma^2), \quad |\phi| < 1.$$

- If we don't believe in normality of ε_t , we have quasi-MLE.
- If we do: MLE.

If $\varepsilon \sim N$ so is $Y_t|Y_{t-1} \sim N$:

$$Y_t|Y_{t-1} \sim N(c + \phi Y_{t-1}, \sigma^2).$$

Conditional MLE

- If we know Y_{t-1} the only term that is random is ε_t.

$$f(y_t|y_{t-1},\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t - c - \phi y_{t-1})^2},$$

  where $\theta = (c, \phi, \sigma^2)$.
- What about Y_1?

$$\begin{aligned} Y_1 &\sim N(E[Y_t], var(Y_t)) \\ &\sim N(\mu, \frac{\sigma^2}{1-\phi^2}), \qquad \mu = \frac{c}{1-\phi}. \end{aligned}$$

- If $\phi = 1$ no unconditional mean or variance exist.

Exact MLE

Maximum likelihood

$$L(\tilde{\theta}|y_1,\ldots,y_T) = f(y_1|\tilde{\theta}) \times \prod_{t=2}^{T} f(y_t|y_{t-1},\tilde{\theta}),$$

i.e.

$$L(\tilde{\theta}|y_1,\ldots,y_T) = \frac{1}{\sqrt{2\pi(\frac{\sigma^2}{1-\phi^2})}} e^{-\frac{1}{2\sigma^2/(1-\phi^2)}(y_1-\frac{c}{1-\phi})^2} \times \prod_{t=2}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_t-c-\phi y_{t-1})^2}$$

- Need to solve non-linear optimization problem.

MA(1)

Recall

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \qquad \varepsilon_t \sim iidN(0, \sigma^2), \ |\theta| < 1.$$

$|\theta| < 1$ *is assumed for invertible representation only.*

Figure: Bi-model likelihood function for MA process

Estimation MA(1)

$$\begin{aligned} Y_t|\varepsilon_{t-1} &\sim N(\mu+\theta\varepsilon_{t-1},\sigma^2), \\ f(y_t|\varepsilon_{t-1},\tilde{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_t-\mu-\theta\varepsilon_{t-1})^2}, \end{aligned}$$

$\tilde{\theta} = (\mu, \theta, \sigma^2)$.

- *Problem: without knowing* ε_{t-2} *we don't observe* ε_{t-1}. *Need to know* ε_{t-2} *to compute* $\varepsilon_{t-1} = y_t - \mu - \theta\varepsilon_{t-2}$.
- But ε_{t-2} unobservable.
- Assume $\varepsilon_0 = 0$.
- *Make it non-random, just fix it with number 0. The trick works with any number.*

Estimation MA(1)

$$\begin{aligned} Y_1|\varepsilon_0 &\sim N(\mu, \sigma^2), \\ Y_1 &= \mu + \varepsilon_1 \Rightarrow \varepsilon_1 = Y_1 - \mu, \\ Y_2 &= \mu + \varepsilon_2 + \theta\varepsilon_1 \Rightarrow \varepsilon_2 = Y_2 - \mu - \theta(Y_1 - \mu), \\ \varepsilon_t &= Y_t - \mu - \theta(Y_{t-1} - \mu) + \ldots + (-1)^{t-1}\theta^{t-1}(Y_1 - \mu). \end{aligned}$$

- Conditional likelihood ($\varepsilon_0 = 0$).:

$$L(\tilde{\theta}|y_1, \ldots, y_T, \varepsilon_0 = 0) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\varepsilon_t^2}.$$

- If $|\theta| < 1$ *(much less)*, ε_0 doesn't matter, CMLE is consistent.
- Exact MLE requires Kalman Filter.

FORECASTING

Simple Model

- With estimated parameters of the model we can make a forecast about the future behavior of the variable of interest.
- Simple Model:

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim WN,$$

- *Today's GDP growth depends on yesterday's GDP growth,*
- *Relationship given, for example, by economic theory.*
- *Once agents know the structure of the model they can form a forecast.*

Notation

Denote:

$\{Y_t\}$ – covariance-stationary process, e.g. $ARMA(p,q)$,

Ω_t – information available at time t,

$Y^*_{t+1|t}$ – forecast of Y_{t+1} based on Ω_t.

- *In our simple model it is Y_{t+1} given Y_t.*

Loss Function

- *We want to assess how good the forecast is.*
- *Assume people don't like mistakes in any direction.*

Rule: evaluate forecast with a quadratic loss function:

$$\min E[(Y_{t+1} - Y^*_{t+1|t})^2] = MSE(Y_{t+1}, Y^*_{t+1|t}).$$

Result: The minimum MSE forecast of Y_{t+1} based on Ω_t is $E[Y_{t+1}|\Omega_t]$.

- *Y_{t+1} is modeled as a random variable (have the whole distribution) but it's behavior is summarized by its mean.*

Linear Projection

- *Y_{t+1} may be very complicated and calculating $E[Y_{t+1}|\Omega_t]$ might be very cumbersome.*
- If $E[Y_{t+1}|\Omega_t]$ difficult to compute, use "best linear forecast".

  $X_{t_{p\times 1}}$ – variables in Ω_t "useful" for prediction
- Linear projection:

$$\hat{Y}_{t+1|t} = \alpha' X_t = \alpha_1 X_{1t} + \ldots + \alpha_p X_{pt},$$

where

$$E[(Y_{t+1} - \alpha' X_t) \cdot X_{it}] = 0, \qquad i = 1, \ldots, p.$$

  - *p moments conditions ensure that error is orthogonal to any information in Ω_t:*
  - *forecast errors are uncorrelated with past information.*

MSE Linear Forecast

Result: The minimum MSE linear forecast of Y_{t+1} is a linear projection.

- For Gaussian (Normal) process: $E[Y_{t+1}|\Omega_t] = \hat{Y}_{t+1|t}$.
- *Linear projection is optimal in Gaussian case.*
- *How do we deal with αs ?*
  - $\hat{Y}_{t+1|t} = \alpha' X_t$ can be thought of as computed by OLS,
  - If $\{X_t, Y_t\}$ is covariance stationary and ergodic, $b \xrightarrow{p} \alpha$, i.e. OLS estimate, b, converges in probability to the true value, α.
- *Optimality is defined in terms of quadratic loss function.*

ARMA Models

Solve Wold form

$$
\begin{aligned}
Y_t - \mu &= \psi(L)\varepsilon_t, \qquad \varepsilon_t \sim WN \\
&\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j, \quad \psi_0 = 1, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty \\
Y_{t+s} &= \mu + \varepsilon_{t+s} + \psi_1 \varepsilon_{t+s-1} + \ldots + \psi_s \varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \ldots, \\
\hat{Y}_{t+1|t} &= \mu + \psi_s \varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \ldots.
\end{aligned}
$$

- *the last line uses information available at time t and* $E_t[\varepsilon_{t+i}] = 0, \ i > 0.$

ARMA Models

$$\begin{aligned} MSE(\hat{Y}_{t+s|t}, Y_{t+s}) &= E[(\varepsilon_t + \psi\varepsilon_{t+s-1} + \ldots + \psi_{s-1}\varepsilon_{t+1})^2] \\ &= \sigma^2(1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{s-1}^2) \leq var(Y_{t+s}). \end{aligned}$$

- *We are better off with linear projection than with unconditional variance.*

But,

$$\lim_{s\to\infty} \sigma^2 \sum_{k=0}^{s} \psi_k^2 = var(Y_t).$$

- *Upper limit for uncertainty is as high as the unconditional variance.*

Forecasting ARMA Models: AR(1)

E.g. AR(1):

$$Y_t - \mu = \phi(Y_{t-1} - \mu) + \varepsilon_t$$

One period ahead forecast:

$$\hat{Y}_{t+1|t} = \mu + \phi(Y_t - \mu)$$

Using Wold formulation:

$$\begin{aligned}
\psi_j &= \phi^j, \\
\Rightarrow \hat{Y}_{t+s|t} &= \mu + \phi^s \varepsilon_t + \phi^{s+1} \varepsilon_{t-1} + \ldots \\
&= \mu + \phi^s (\varepsilon_t + \phi \varepsilon_{t-1} + \ldots) \\
&= \mu + \phi^s (Y_t - \mu), \\
\lim_{s \to \infty} MSE &= \sigma^2 \frac{1}{1 - \phi^2} = var(Y_t).
\end{aligned}$$

Forecasting ARMA Models: MA(1)

E.g. MA(1):

$$\psi_0 = 1, \quad \psi_1 = \theta, \quad , \psi_j = 0, \quad \forall j > 1,$$

$$\begin{aligned} \hat{Y}_{t+1|t} &= \mu + \theta\hat{\varepsilon}_t, \qquad \hat{\varepsilon}_t = (Y_t - \mu) - \theta\hat{\varepsilon}_{t-1}, \\ \hat{Y}_{t+s|t} &= \mu, \qquad \forall s > 1 \\ \lim_{s\to\infty} MSE &= \sigma^2(1+\theta^2) = var(Y_t). \end{aligned}$$

Forecast error is just a deviation of the series from the long-run unconditional mean.

Forecasting ARMA Models: AR(2)

E.g. AR(2):

$$\underbrace{\begin{bmatrix} Y_t - \mu \\ Y_{t-1} - \mu \end{bmatrix}}_{\beta_t} = \underbrace{\begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} Y_{t-1} - \mu \\ Y_{t-2} - \mu \end{bmatrix}}_{\beta_{t-1}} + \underbrace{\begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}}_{v_t}$$

$$\beta_t = F\beta_{t-1} + v_t$$

Then,

$$\begin{aligned} \hat{\beta}_{t+s|t} &= F^s v_t + F^{s+1} v_{t-1} + \dots \\ &= F^s (v_t + F v_{t-1} + \dots) = F^s \beta_t. \end{aligned}$$

Let

$$F^s = \begin{bmatrix} f_{11}^{(s)} & f_{12}^{(s)} \\ f_{21}^{(s)} & f_{22}^{(s)}. \end{bmatrix}$$

Then,

$$\hat{Y}_{t+s|t} = \mu + f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu)$$

Finite-sample forecast

- Forecasts based on Wold form assume infinite number of observations.
- *We don't have them in reality.*
- For finite number of observation:
  - use approximation: set presample $\varepsilon_\tau = 0$,
  - do exact finite-sample forecast.
- Kalman filter calculates linear projections for finite number of observations,
  - exact finite-sample forecast,
  - allow for exact MLE of ARMA models based on prediction error decomposition.
- See Hamilton chapter 4 for alternative.

Forecast accuracy

- $\{Y_t\}$ is the series to be forecast.
- $\{Y^1_{t+h|t}\}$ and $\{Y^2_{t+h|t}\}$ are two competing forecasts of Y_{t+h} based on Ω_t. *(for example, from an AR(p) and ARMA(p,q) models, respectively)*
- Forecast errors from the two models are

$$\begin{aligned} \varepsilon^1_{t+h|t} &= y_{t+h} - y^1_{t+h|t}, \\ \varepsilon^2_{t+h|t} &= y_{t+h} - y^2_{t+h|t}, \end{aligned}$$

  producing series of serially correlated forecasts $\{\varepsilon^1_{t+h|t}\}^T_{t_0}$, $\{\varepsilon^2_{t+h|t}\}^T_{t_0}$,
  - h-step forecasts use overlapping data.
- Forecast accuracy can be measured by a loss function

$$L(y_{t+h}, y^i_{t+h|t}) = L(\varepsilon^i_{t+h|t})$$

  - squared error loss: $L(\varepsilon^i_{t+h|t}) = (\varepsilon^i_{t+h|t})^2$,
  - absolute error loss: $L(\varepsilon^i_{t+h|t}) = \left|\varepsilon^i_{t+h|t}\right|$.

H_0

- To test if one model predicts better than another consider null hypothesis

$$H_0: \quad E\left(L(\varepsilon^1_{t+h|t})\right) = E\left(L(\varepsilon^2_{t+h|t})\right)$$

against

$$H_1: \quad E\left(L(\varepsilon^1_{t+h|t})\right) \neq E\left(L(\varepsilon^2_{t+h|t})\right)$$

- Define loss deferential

$$d_t = L(\varepsilon^1_{t+h|t}) - L(\varepsilon^2_{t+h|t}).$$

- The null hypothesis of equal predictive accuracy is

$$H_0: \quad E(d_t) = 0.$$

Diebold-Mariano test

- The Diebold-Mariano test statistics is

$$S = \frac{\bar{d}}{\left(\widehat{avar}(\bar{d})\right)^{1/2}} = \frac{\bar{d}}{\left(\widehat{LRV}_{\bar{d}}/T\right)^{1/2}}$$

where

$$\begin{aligned} \bar{d} &= \frac{1}{T_0}\sum_{t=t_0}^{T} d_t \\ LRV_{\bar{d}} &= \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j, \quad \gamma_j = cov(d_t, d_{t-j}) \end{aligned}$$

and $\widehat{LRV}_{\bar{d}}$ is a consistent estimate of asymptotic (long-run) variance of $\sqrt{T}\bar{d}$

- Diebold-Mariano (1995) show that under the null of equal predictive accuracy

$$S \overset{A}{\sim} N(0,1).$$

  - Reject the null of equal predictive accuracy at the 5% level if $|S| > 1.96$.
  - One sided tests may also be computed.

Prediction Error Decomposition

Likelihood

- Consider $\{Y_t\}$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} = \tilde{y}_T \sim N(\mu_{T\times 1}, \Omega_{T\times T}),$$

- Since it is covariance stationary process, each Y_t has the same mean and variance, $\omega_{11} = \sigma^2 = \omega_{22} = \omega_{TT}$.

$$\Omega = \begin{bmatrix} \omega_{11} & \omega_{12} & & \omega_{1T} \\ \omega_{21} & \omega_{22} & & \vdots \\ \vdots & & \ddots & \\ \omega_{T1} & \dots & & \omega_{TT} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & & \vdots \\ \vdots & & \ddots & \\ \gamma_{T-1} & \dots & & \gamma_0 \end{bmatrix}, \text{ as } \begin{array}{r} \omega_{jj} = \gamma_0, \\ \omega_{ij} = \gamma_{i-j}, \\ i > j. \end{array}$$

- The likelihood function:

$$L(\tilde{\theta}|\tilde{y}_T) = (2\pi)^{-\frac{T}{2}} \det(\Omega)^{-\frac{1}{2}} e^{-\frac{1}{2}(\tilde{y}_T-\mu)'\Omega^{-1}(\tilde{y}_T-\mu)}.$$

Factorization

- For large T, Ω might be large and difficult to invert.
- Since Ω is positive definite symmetric matrix, there exists a unique, triangular factorization of Ω,

$$\Omega = AfA',$$

where

$$f_{T\times T} = \begin{bmatrix} f_1 & 0 & \cdots & 0 \\ 0 & f_2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & f_T \end{bmatrix}, \qquad f_t > 0 \ \forall t \quad \textit{diagonal matrix}$$

$$A_{T\times T} = \begin{bmatrix} 1 & & & 0 \\ a_{21} & 1 & & \\ \vdots & & \ddots & \\ a_{T1} & a_{T2} & \ldots & 1 \end{bmatrix}, \qquad \textit{lower triangular matrix}$$

Prediction errors

- Define $\eta = A^{-1}(\tilde{y}_T - \mu)$,

$$E(\eta) = 0, \quad var(\eta) = var(A^{-1}(\tilde{y}_T - \mu)) = A^{-1}\Omega[A']^{-1} = A^{-1}AfA'[A']^{-1} = f.$$

- Since f is diagonal, η is a series of random variables that are uncorrelated with each other, i.e $E(\eta_t\eta_\tau) = 0, \quad , \forall t \neq \tau$.
- Since $A\eta = (\tilde{y}_T - \mu)$ and A is a lower-triangular matrix with $1s$ along the principal diagonal,

$$\begin{aligned}
\eta_1 &= y_1 - \mu \\
\eta_2 &= y_2 - \mu - a^*_{11}\eta_1 \\
\eta_3 &= y_3 - \mu - a^*_{21}\eta_1 - a^*_{22}\eta_2 \\
&\vdots \\
\eta_T &= y_T - \mu - \sum_{i=1}^{T-1} a^*_{T-1,i}\eta_i
\end{aligned}$$

where $a^*_{i,j} = A_{i+1,j}$

Prediction errors

- Since $E(\eta_t \eta_\tau) = 0$,

$$E(\eta_2 \eta_1) = E\left((y_2 - \mu - a_{11}^*(y_1 - \mu))(y_1 - \mu)\right) = 0$$

  so the a_{11}^* is the coefficient of the linear projection of $(y_2 - \mu)$ on $(y_1 - \mu)$ and $\eta_2 = y_2 - y_{2|1}$ is a prediction error .
- Similarly, $E(\eta_3 \eta_2) = E(\eta_3 \eta_1) = 0$ imply that a_{21}^* and a_{22}^* are linear projection coefficients of $(y_3 - \mu)$ on $(y_2 - \mu)$ and $(y_2 - \mu)$, with $\eta_3 = y_3 - y_{3|2}$
- Therefore, η_t is t^{th} element of $\eta_{T \times 1}$ = prediction error $y_t - \hat{y}_{t|t-1}$.

Likelihood

- The likelihood function can be rewritten as:

$$L(\tilde{\theta}|\tilde{y}_T) = (2\pi)^{-\frac{T}{2}} \det(AfA')^{-\frac{1}{2}} e^{-\frac{1}{2}(\tilde{y}_T-\mu)'(AfA')^{-1}(\tilde{y}_T-\mu)}$$

- As A is lower triangular with $1s$ along the principal diagonal, $\det(A) = 1$ and

$$\det(AfA) = \det(A) \cdot \det(f) \cdot \det(A') = \det(f).$$

- Then,

$$\begin{aligned} L(\tilde{\theta}|\tilde{y}_T) &= (2\pi)^{-\frac{T}{2}} \det(f^{-1})^{-\frac{1}{2}} e^{-\frac{1}{2}\eta'(f^{-1})^{-1}\eta} \\ &= \prod_{t=1}^{T} \left(\frac{1}{\sqrt{2\pi f_t}} e^{-\frac{1}{2}\frac{\eta_t^2}{f_t}} \right), \end{aligned}$$

where η_t is t^{th} element of $\eta_{T\times 1}$ = prediction error $y_t - \hat{y}_{t|t-1}$,

$$\hat{y}_{t|t-1} = \sum_{i=1}^{t-1} a_{t,i}^* y_i, \qquad i = 2, 3, \ldots, T.$$

Kalman Filter

Note: Given $y_t \sim N(\mu, \Omega)$,

$$\eta_t | \Omega_{t-1} \sim N(0, f_t),$$

where f_t is an (t,t) diagonal element of f matrix,

$$\ln L = -\frac{1}{2}\sum_{t=1}^{T} \ln(2\pi f_t) - \frac{1}{2}\sum_{t=1}^{T} \frac{\eta_t^2}{f_t},$$

since $\eta_t \sim N$ and independent of each other.

- The Kalman filter recursively calculates linear projection of y_t on past information Ω_{t-1} for any model that can be cast in state-space form.
- *Kalman filter: for any structure it solves for linear prediction.*

STATE-SPACE

Measurement (Observation) Equation

General form that encompasses a wide variety od models.

1. Measurement (Observation) Equation
   - *Represent the static relationship between observed variables (data) and unobserved state variables.*

$$y_t = H_t \beta_t + A z_t + e_t,$$

where y_t denotes observed data, β_t is a state vector that captures the dynamics, z_t is exogenous, observed variables *for example, lagged values of* y_t *but also other data*, and e_t is an error term,

$$e_t \sim N(0, R).$$

The existence of the state vector makes this representation not a simple linear model.

Transition (State) Equation

2. Transition (State) Equation
   - *Captures the dynamics in the system, causes the system to go on and on.*

$$\beta_t = \tilde{\mu} + F\beta_{t-1} + v_t,$$

where $\tilde{\mu}$ is a vector of constants, F is the transition matrix, and v_t is an error vector,

$$v_t \sim N(0, Q).$$

   - *Like AR(1) but in vector/matrix form.*

Transition (State) Equation

$$\beta_t = \tilde{\mu} + \boldsymbol{F}\beta_{t-1} + v_t,$$

- The state vector has and AR(1) kind of representation.
- Describes evolution of state vector.
- These state vectors can be unobservable.
- Transition equation can be used to get information about the unobservable, conditioning on data which is observable (Bayesian).

Error terms

Error terms:

$$e_t \sim N(0, R), \quad v_t \sim N(0, Q),$$

where R, Q are var-cov matrices and

$$E[e_t v_\tau'] = 0, \quad \forall t, \tau$$

- Restrictive assumption
- The model can be represented in a way that is not very restrictive.
- Even with $E[e_t v_\tau'] \neq 0$ we can estimate the model with (modified) Kalman Filter but it becomes more complicated.
- The normality assumption might not be always good...

  ...but it allows to use MLE.

Examples: AR(2)

Consider an AR(2) process

$$\begin{aligned} y_t &= c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \\ \varepsilon_t &= WN(0, \sigma^2). \end{aligned}$$

State equation

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}$$

Observation equation

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix}.$$

Examples: AR(2) again

Consider again an AR(2) process

$$\begin{aligned} y_t &= c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t, \\ \varepsilon_t &= WN(0, \sigma^2). \end{aligned}$$

State equation

$$\beta_t = \begin{bmatrix} y_t \\ \phi_2 y_{t-1} \end{bmatrix} = \begin{bmatrix} \phi_1 & 1 \\ \phi_2 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ \phi_2 y_{t-2} \end{bmatrix} + \begin{bmatrix} c \\ 0 \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}$$

Observation equation

$$y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} y_t \\ \phi_2 y_{t-1} \end{bmatrix}.$$

Examples: MA(1)

- Consider an MA(1) process

$$\begin{aligned} y_t &= \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \\ \varepsilon_t &= WN(0, \sigma^2). \end{aligned}$$

- Define

$$\beta_t = \begin{bmatrix} y_t - \mu \\ \theta\varepsilon_t \end{bmatrix}$$

- Then

State equation

$$\beta_t = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \beta_{t-1} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} \varepsilon_t$$

Observation equation

$$y_t = \mu + \begin{bmatrix} 1 & 0 \end{bmatrix} \beta_t.$$

Examples: ARMA(1,1)

ARMA(1,1):
Set $\mu = 0$,

$$y_t = \phi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}, \qquad \varepsilon_t \sim N(0, \sigma^2).$$

- *There might be more than one way to represent a model in a state-space form.*
- *There might be differences in efficiency between different ways.*

Examples: ARMA(1,1)

State equation:

- The general form
$$\beta_t = F\beta_{t-1} + v_t.$$
- Put
$$\beta_t = \begin{bmatrix} y_t \\ \varepsilon_t \end{bmatrix} \Rightarrow \beta_{t-1} = \begin{bmatrix} y_{t-1} \\ \varepsilon_{t-1} \end{bmatrix}$$
- Put $y_t = \phi y_{t-1} + \theta\varepsilon_{t-1} + \varepsilon_t$ in a matrix notation:
$$\underset{\beta_t}{\begin{bmatrix} y_t \\ \varepsilon_t \end{bmatrix}} = \underset{F}{\begin{bmatrix} \phi & \theta \\ 0 & 0 \end{bmatrix}} \underset{\beta_{t-1}}{\begin{bmatrix} y_{t-1} \\ \varepsilon_{t-1} \end{bmatrix}} + \underset{v_t}{\begin{bmatrix} \varepsilon_t \\ \varepsilon_t \end{bmatrix}},$$

and $v_t \sim N(0, Q)$, $Q = \begin{bmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{bmatrix}$.

y_t – *observable*, ε_t–*unobservable, forecast error*

Examples: ARMA(1,1)

Observation equations:

$$\begin{array}{ccccc} y_t & = & \begin{bmatrix} 1 & 0 \end{bmatrix} & \begin{bmatrix} y_t \\ \varepsilon_t \end{bmatrix} \\ y_t & & H & \beta_t \end{array}$$

no exogenous variables: $A = 0$, also $R = 0$.

$y_t = H\beta_t$ for this case (ARMA(1,1)).

The parameters ϕ, θ, σ^2 are captured in F, Q matrices. The Kalman Filter will estimate them.

- *For KF what goes in β_t doesn't matter.*
- *Only parameters F, Q, H, R will matter.*
- *The state vector is now defined by F, Q, H, and the observations.*

ARMA(1,1): Alternative Representation

A more "elegant" (i.e. easier for computation) representation.
Lag notation (alternative representation for ARMA(1,1))

$$\begin{aligned}(1-\phi L)y_t &= (1+\theta L)\varepsilon_t \\ y_t &= (1-\phi L)^{-1}(1+\theta L)\varepsilon_t \\ y_t &= (1+\theta L)(1-\phi L)^{-1}\varepsilon_t.\end{aligned}$$

Define $x_t = (1-\phi L)^{-1}\varepsilon_t$

$$\begin{aligned}(1-\phi L)x_t &= \varepsilon_t, \qquad (x_t \text{ is AR(1), } \textit{not observed}) \\ x_t - \phi x_{t-1} &= \varepsilon_t\end{aligned}$$

Then,

$$\begin{aligned}y_t &= (1+\theta L)x_t \\ y_t &= x_t + \theta x_{t-1}.\end{aligned}$$

So y_t is a linear combination of 2 unobservable AR(1) processes, x_t and x_{t-1}.

ARMA(1,1): State-Space

Observation equation *(all randomness in the state equation)*

$$y_t = H\beta_t,$$

where

$$y_t = \begin{bmatrix} 1 & \theta \end{bmatrix} \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}$$

- *Inside H there are parameters to be estimated.*
- *$A = 0$, no exogenous, $R = 0$ as the observable equation is just the identity (no randomness of e_t).*

ARMA(1,1): State-Space

State equation

$$\underbrace{\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}}_{\beta_t} = \underbrace{\begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} x_{t-1} \\ x_{t-2} \end{bmatrix}}_{\beta_{t-1}} + \underbrace{\begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}}_{v_t},$$

so

$$v_t \sim N(0, Q), \qquad Q = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

- *So ϕ is in F, θ in H, and σ^2 in Q.*

Given F, Q, H, A, R and data (y_t's), use Kalman Filter to find prediction error decomposition of joint likelihood for $\tilde{y}_T = (y_1, \ldots y_T)$, given by $L(\theta, \phi, \sigma^2 | \tilde{y}_T)$. *(exact likelihood)*

KALMAN FILTER

Kalman Filter

Kalman filter:

- *purpose: to make inference about unobservable given the observable,*
- *application: signal extraction in engineering,*
- *economics: don't know the parameters F, Q, H and want to estimate them.*

State-space form

ME: Measurement (Observation) equation:

$$y_t = H\beta_t + e_t, \qquad e_t \sim N(0, R)$$

SE: Transition (State) equation:

$$\beta_t = \tilde{\mu} + F\beta_{t-1} + v_t, \qquad v_t \sim N(0, Q),$$

$$E[e_t v_\tau'] = 0.$$

Mean of β

1. β_t is a random variable
   - it might be unobservable and no data for it,
   - it is normal random variable as it is sum of normal variables, $v_t \sim N$.

Conditional mean

$$\begin{aligned} \beta_t|\Omega_{t-1} &\sim N\left(E[\beta_t|\Omega_{t-1}], var(\beta_t|\Omega_{t-1})\right) \\ E[\beta_t|\Omega_{t-1}] &= \beta_{t|t-1}, \qquad \textit{conditional expectations.} \end{aligned}$$

- *We may not know what β's are.*
- *If we have information about its distribution, we can calculate mean, variance, etc.*
- *β_{t-1} may be not observable: take expectations of it*

$$\begin{aligned} E[\beta_t|\Omega_{t-1}] \equiv \beta_{t|t-1} &= \tilde{\mu} + FE[\beta_{t-1}|\Omega_{t-1}] + 0 \\ \beta_{t|t-1} &= \tilde{\mu} + F\beta_{t-1|t-1}, \end{aligned}$$

- *In AR(1): $E[y_t] = \mu + \phi E[y_{t-1}]$, last term is observable.*

Variance of β

Conditional variance

$$Var(\beta_t|\Omega_{t-1}) \equiv P_{t|t-1} = E[(\beta_t - \beta_{t|t-1})(\beta_t - \beta_{t|t-1})'].$$

Recall

$$var(ax) = a^2 var(x), \qquad a - scalar, x - random\ vector.$$

Two sources of randomness (variation) for β_t:

1. v_t is a random variable,
2. β_{t-1} is also random so there might be difference between β_{t-1} and $\beta_{t|t-1}$, *there may not be equal to each other.*

$$P_{t|t-1} = F P_{t-1|t-1} F' + Q,$$

where $P_{t|t-1}$, uncertainty about β_t equals sum of uncertainty about β_{t-1}, $P_{t-1|t-1}$, and uncertainty about v_t.
Note: $cov(\beta_{t-1}, v_t) = 0$.

y_t

2. y_t is a random variable.
   - Now, we have data on y_t.
   - We have some joint density of y_t, β_t and some prior.
   - Using data we get posterior of β_t.

   - *We want to make inference for β_t which we don't observe.*
   - *We see y_t which is related to β_t.*
   - *We make inferences on β_t by observing joint density (distribution) of ys and βs (Bayesian view).*

Distribution of y_t

Distribution of y_t given state-space

$$y_t|\Omega_{t-1} \sim N(E[y_t|\Omega_{t-1}], var(y_t|\Omega_{t-1}),$$

- Conditional mean

$$E[y_t|\Omega_{t-1}] \equiv y_{t|t-1} = H\beta_{t|t-1} + 0$$

- Conditional variance

$$var(y_t|\Omega_{t-1}) \equiv f_{t|t-1} = HP_{t|t-1}H' + R,$$

since we don't know β_t.

- Note: $cov(H\beta_t, e_t) = 0$ because $E[v_t e_t] = 0$.

If $E[v_t e_t] \neq 0$ *we will add another term in the* $var(y_t|\Omega_{t-1})$ *capturing that.*

Joint Distribution

- Covariance between β_t and y_t:

$$cov(y_t, \beta_t | \Omega_{t-1}) = P_{t|t-1} H',$$

$$\text{as } cov(H\beta_t + e_t, \beta_t) = cov(H\beta_t, \beta_t) + cov(e_t, \beta_t) = cov(\beta_t, \beta_t)H' + 0.$$

Then, the joint distribution for y_t and β_t is joint normal:

$$\begin{bmatrix} \beta_t \\ y_t \end{bmatrix} \Big| \Omega_{t-1} \sim N\left(\begin{bmatrix} \beta_{t|t-1} \\ H\beta_{t|t-1} \end{bmatrix}, \begin{bmatrix} P_{t|t-1} & P_{t|t-1}H' \\ P_{t|t-1}H' & f_{t|t-1} \end{bmatrix} \right).$$

Kalman Filter

Two steps of Kalman Filter :

(a) Prediction,

(b) Given y_t updating inference on β_t.

Definition

Given $\beta_{0|0}$, $P_{0|0}$, Kalman Filter solves the following six equations for $i = 1, \ldots, T$

Prediction of y_t, β_t

$$(1) \quad \beta_{t|t-1} = \tilde{\mu} + F\beta_{t-1|t-1},$$

$$(2) \quad P_{t|t-1} = F P_{t-1|t-1} F' + Q,$$

Forecast error:

$$(3) \quad \eta_{t|t-1} \equiv y_t - y_{t|t-1} = y_t - H\beta_{t|t-1},$$

Variance of forecast error:

$$(4) \quad f_{t|t-1} = H P_{t|t-1} H' + R$$

Updating of y_t, β_t

$$(5) \quad \beta_{t|t} = \beta_{t|t-1} + \kappa_t \eta_{t|t-1},$$

$$(6) \quad P_{t|t} = P_{t|t-1} - \kappa_t H P_{t|t-1},$$

$$\kappa_t \equiv P_{t|t-1} H' f_{t|t-1}^{-1} \quad \text{“Kalman gain”}.$$

Kalman Filter

- $\beta_{0|0}$, $P_{0|0}$, are equal to unconditional mean and variance, and reflect prior beliefs.
- If the state space model is covariance stationary,

$$\begin{aligned} E[\beta] = \beta_{0|0} &= (\mathbb{I} - \mu)^{-1}\tilde{\mu} \\ var(\beta) = P_{0|0} &= FP_{0|0}F' + Q \\ vec(P_{0|0}) &= vec(FP_{0|0}F') + vec(Q) \\ vec(P_{0|0}) &= (F \otimes F)vec(P_{0|0}) + vec(Q) \\ vec(P_{0|0}) &= (\mathbb{I} - F \otimes F)^{-1}vec(Q). \end{aligned}$$

  since $vec(ABC) = (C' \otimes A)vec(B)$.
- Equation (5) is a linear combination of previous guess and forecast error.

$$\begin{aligned} (5) \quad \beta_{t|t} &= \beta_{t|t-1} + \kappa_t \eta_{t|t-1}, \\ (6) \quad P_{t|t} &= P_{t|t-1} - \kappa_t H P_{t|t-1}, \\ \kappa_t &\equiv P_{t|t-1}H' f_{t|t-1}^{-1} \quad \text{“Kalman gain”}. \end{aligned}$$

Kalman Gain

- The stronger the covariance between y_t and β_t, the more we will update when we see high forecast error.
- If the relationship is weaker, we don't put much weight as probably it is not driven by β_t.
- The weight depends on the variance of forecast error: if f^{-1} big, put high weight on that observations.
- Once we have $\eta_{t|t-1}, f_{t|t-1}$, we can do MLE after constructing the joint likelihood of prediction error decomposition.
  - *The Kalman gain depends on the relationship between y_t and β_t since $P_{t|t-1}H' = cov(\beta_t, y_t)$ and $f_{t|t-1}^{-1}$ is the precision of the forecast error.*
  - *The bigger the variance of forecast error the smaller the Kalman gain and less weight put to updating.*
  - *Equation (6) measures conditional variance.*
  - *Since we observe y_t the uncertainty declines.*